Salud, Ciencia y Tecnología. 2025; 5:2320
doi: 10.56294/saludcyt20252320

ORIGINAL

Hybrid CNN-Capsule Network Architecture for Automated Diabetic Retinopathy Classification: A Rigorous Statistical Validation on Clinical Data

Arquitectura Híbrida de Red CNN-Cápsula para la Clasificación Automatizada de la Retinopatía Diabética: Una Validación Estadística Rigurosa en Datos Clínicos

Mini T.V[1]

[1]Associate Professor, Department of Computer Science, Sacred Heart College (Autonomous). Chalakudy, Kerala, India.

Cite as: T. V M. Hybrid CNN-Capsule Network Architecture for Automated Diabetic Retinopathy Classification: A Rigorous Statistical Validation on Clinical Data. Salud, Ciencia y Tecnología. 2025; 5:2320. https://doi.org/10.56294/saludcyt20252320

Submitted: 03-07-2025 **Revised:** 08-09-2025 **Accepted:** 11-11-2025 **Published:** 12-11-2025

Editor: Prof. Dr. William Castillo-González

Corresponding Author: Mini T. V

ABSTRACT

Introduction: diabetic retinopathy (DR) remains the leading cause of preventable blindness worldwide, necessitating the development of automated screening solutions to address the global shortage of ophthalmological expertise. Many existing deep learning approaches lack rigorous statistical validation and fail to address the severe class imbalance characteristic of clinical DR populations.
Objective: a hybrid convolutional neural network-capsule network architecture for automated DR severity grading, emphasizing statistical rigor and clinical applicability over benchmark optimization.
Method: a novel ensemble architecture combining ResNet-50 and MobileNet-V2 feature extractors with capsule network classifiers was developed and rigorously validated using 5-fold stratified cross-validation on the Messidor dataset (n=1200) with strict patient-level data splitting. The framework addresses severe class imbalance through systematic multi-level mitigation strategies including SMOTE-ENN augmentation, focal loss optimization (α=0,25, γ=2,0), and dynamic class weighting. Statistical robustness was ensured through bootstrap confidence intervals (n=1000 iterations), McNemar's paired comparison tests, and comprehensive ablation studies.
Results: the hybrid architecture achieved 88,67 % ± 2,43 % overall accuracy (95 % CI: 86,24 %-91,10 %), with an area under the curve of 90,85 % ± 2,15 % (95 % CI: 88,70 %-93,00 %). Critically, sensitivity for sight-threatening cases reached 84,0 % for severe NPDR and 80,6 % for proliferative DR, while maintaining 91,8 % specificity for non-referrable cases. The systematic class imbalance mitigation strategy improved the F1-scores of the minority class by 19,3 % relative to the standard CNN baseline ($p < 0,001$, McNemar's $x^2 = 15,67$). Cross-validation consistency (coefficient of variation: 2,74 %) demonstrated model stability essential for clinical deployment.
Conclusions: the hybrid CNN-capsule architecture provides clinically relevant DR classification with transparent statistical validation. The demonstrated sensitivity for sight-threatening cases and conservative classification patterns support potential screening application, though prospective clinical validation remains necessary.

Keywords: Diabetic Retinopathy; Computer-Aided Diagnosis; Capsule Networks; Deep Learning; Medical Image Analysis; Clinical Validation.

RESUMEN

Introducción: la retinopatía diabética (RD) sigue siendo la principal causa de ceguera prevenible a nivel mundial, lo que hace necesaria la creación de soluciones automatizadas de cribado ante la escasez global de especialistas en oftalmología. Muchos enfoques actuales basados en aprendizaje profundo carecen de una validación estadística rigurosa y no abordan adecuadamente el grave desequilibrio de clases característico de las poblaciones clínicas con RD.

Objetivo: desarrollar una arquitectura híbrida de red neuronal convolucional y red de cápsulas para la clasificación automatizada de la gravedad de la RD, con énfasis en la rigurosidad estadística y la aplicabilidad clínica, por encima de la mera optimización de métricas de referencia.
Método: se desarrolló una arquitectura de ensamble novedosa que combina extractores de características ResNet-50 y MobileNet-V2 con clasificadores basados en redes de cápsulas, validada rigurosamente mediante validación cruzada estratificada de 5 pliegues sobre el conjunto de datos Messidor (n=1200), con una estricta división de datos a nivel de paciente. El marco aborda el severo desequilibrio de clases mediante estrategias de mitigación multinivel, incluyendo aumento de datos con SMOTE-ENN, optimización con función de pérdida focal (α=0,25, γ=2,0) y ponderación dinámica de clases. La solidez estadística se garantizó mediante intervalos de confianza generados por bootstrap (n=1000 iteraciones), pruebas apareadas de McNemar y estudios de ablación exhaustivos.
Resultados: la arquitectura híbrida alcanzó una precisión global de 88,67 % ± 2,43 % (IC del 95 %: 86,24 %-91,10 %), con un área bajo la curva de 90,85 % ± 2,15 % (IC del 95 %: 88,70 %-93,00 %). De manera crítica, la sensibilidad para los casos con riesgo de pérdida de visión alcanzó el 84,0 % para RD no proliferativa severa y el 80,6 % para RD proliferativa, manteniendo una especificidad del 91,8 % para casos no referibles. La estrategia sistemática de mitigación del desequilibrio de clases mejoró las puntuaciones F1 de la clase minoritaria en un 19,3 % con respecto a una red CNN estándar ($p < 0,001$, x^2 de McNemar = 15,67). La consistencia en la validación cruzada (coeficiente de variación: 2,74 %) demostró una estabilidad del modelo esencial para su implementación clínica.
Conclusiones: la arquitectura híbrida CNN–red de cápsulas ofrece una clasificación de la RD clínicamente relevante con una validación estadística transparente. La sensibilidad demostrada para casos con riesgo de pérdida de visión y un patrón de clasificación conservador respaldan su potencial uso en programas de cribado, aunque aún se requiere validación clínica prospectiva.

Palabras clave: Retinopatía Diabética; Diagnóstico Asistido por Computadora; Redes de Cápsulas; Aprendizaje Profundo; Análisis de Imágenes Médicas; Validación Clínica.

INTRODUCTION

Diabetic retinopathy (DR) affects approximately 35 % of the global diabetic population, representing the leading cause of working-age blindness in developed countries.[1] The International Diabetes Federation projects diabetes prevalence to increase from 537 million cases in 2021 to 643 million by 2030, proportionally expanding the population at risk for vision-threatening complications.[2] Early detection and timely intervention can prevent up to 95 % of severe vision loss, yet many healthcare systems lack sufficient ophthalmological resources for comprehensive screening.[3]

Automated DR detection using fundus photography represents a scalable solution to address this screening gap. Deep learning approaches, particularly convolutional neural networks (CNNs), have demonstrated substantial progress in medical image analysis.[4,6] However, critical limitations persist in existing DR detection systems: inadequate statistical validation protocols, insufficient attention to clinical class imbalance, and unrealistic performance claims that hinder clinical translation.[7,8]

Standard CNNs excel at feature extraction but suffer from spatial relationship loss during pooling operations a significant limitation for retinal pathology detection where spatial arrangements of microaneurysms, exudates, and hemorrhages carry diagnostic significance.[9] Capsule networks address this limitation through dynamic routing mechanisms that preserve hierarchical spatial relationships, making them particularly suited for medical imaging applications.[10,11]

This study addresses key methodological gaps in automated DR detection research through three primary contributions:

- Development of a hybrid CNN-capsule architecture that leverages complementary strengths of established feature extractors with spatial-preserving classification
- Implementation of rigorous statistical validation protocols including stratified cross-validation with patient-level splitting, bootstrap confidence intervals, and comprehensive ablation studies,
- Systematic evaluation of class imbalance mitigation strategies with proper baseline comparisons essential for clinical DR populations.

The research prioritizes clinical applicability over benchmark optimization, emphasizing conservative classification bias, minority class sensitivity, and transparent statistical reporting—considerations essential for successful medical AI translation but often overlooked in computer vision research focused primarily on accuracy maximization.

Deep Learning in Diabetic Retinopathy Detection

Gargeya and Leng demonstrated early success in applying deep learning to DR detection, using a dataset of 128 175 images, and achieved 90 % sensitivity and 91 % specificity.[4] However, their work lacked cross-validation and detailed per-class analysis essential for clinical validation. Gulshan et al. developed a system achieving 90,3 % sensitivity and 98,1 % specificity on a large-scale dataset, establishing a foundational benchmark for automated DR detection.[5]

Recent studies have explored various CNN architectures for DR classification. Ting et al. developed a deep learning system for detecting diabetic retinopathy using multi-ethnic retinal images, achieving competitive performance across diverse populations.[12] Abràmoff et al.[13] demonstrated improved automated detection through deep learning integration, emphasizing the importance of rigorous validation protocols.

Advanced Architectures and Validation Methods

Vision transformer approaches have emerged as promising alternatives to traditional CNNs. Dosovitskiy et al.[14] introduced Vision Transformers (ViTs), demonstrating competitive performance on image classification tasks. Meanwhile, Shamshad et al.[15] provided a comprehensive analysis of transformer applications in medical imaging, highlighting their potential for modelling spatial relationships.

Ensemble approaches have shown particular promise in medical applications. Poplin et al.[16] demonstrated that ensemble methods could predict cardiovascular risk factors from retinal photographs, establishing the value of multi-model approaches for medical imaging tasks.

Capsule Networks in Medical Imaging

Capsule networks, introduced by Sabour et al.[17], preserve spatial relationships through dynamic routing between capsule layers. Hinton et al.[18] demonstrated the effectiveness of their approach in object recognition tasks, while recent medical imaging applications have shown promise in preserving anatomical spatial relationships critical for medical diagnosis.

Limited work has explored capsule networks for retinal image analysis specifically. The dynamic routing mechanism's ability to preserve spatial hierarchies makes capsule networks particularly suitable for medical imaging, where anatomical relationships carry diagnostic significance.

Class Imbalance in Medical AI

Clinical DR datasets exhibit severe class imbalance, with sight-threatening cases comprising less than 15 % of typical screening populations.[19] Standard machine learning approaches suffer from majority class bias, potentially missing critical pathological cases. Chawla et al.[20] introduced SMOTE (Synthetic Minority Oversampling Technique), providing a foundation for addressing class imbalance through synthetic sample generation.

Fernández et al.[21] reviewed imbalance learning techniques, emphasizing the importance of evaluation metrics beyond accuracy. Recent advances in focal loss by Lin et al.[22] have shown particular promise for addressing class imbalance in medical imaging applications.

Statistical Validation in Medical AI

Reproducibility and statistical rigor remain significant concerns in medical AI research. Park et al. emphasized the importance of proper validation protocols in medical AI, while Varoquaux and Cheplygina highlighted common statistical pitfalls in machine learning for medical applications.[23,24]

Recent guidelines by Mongan et al.[25] for AI in medical imaging emphasize cross-validation, external validation, and clinical safety analysis—standards that many existing DR detection studies fail to meet comprehensively.

METHOD

Dataset and Experimental Design

This study utilized the Messidor database a publicly available collection of 1,200 color fundus photographs acquired by three ophthalmologic departments using a color video 3CCD camera on a Topcon TRC NW6 non-mydriatic retinograph. Images were captured at 45° field of view with varying resolutions (1440×960, 2240×1488, or 2304×1536 pixels).[26,27] Expert ophthalmologists graded each image according to international clinical diabetic retinopathy severity scales.

The dataset exhibits characteristic clinical class imbalance: No DR (714 images, 59,5 %), Mild non-proliferative DR (178 images, 14,8 %), Moderate NPDR (183 images, 15,2 %), Severe NPDR (94 images, 7,8 %), and Proliferative DR (31 images, 2,6 %). This distribution reflects real-world screening populations where sight-threatening cases constitute a small but clinically critical minority.

Rigorous Validation Protocol

Patient-Level Data Splitting

To prevent data leakage, images were grouped by patient identifier when available, ensuring no patient

appears in both training and validation sets within any fold. For images without explicit patient identifiers, conservative splitting was performed based on acquisition metadata and timestamp clustering.

Stratified Cross-Validation

Five-fold stratified cross-validation was implemented with verification of class distribution maintenance across folds. Chi-square goodness-of-fit tests confirmed no significant deviation from expected proportions (x^2 = 2,14, p = 0,832 across all folds).

Statistical Framework

Bootstrap confidence intervals (n = 1000 iterations) were calculated using the bias-corrected and accelerated (BCa) methodology. McNemar's test assessed statistical significance between model comparisons with explicit reporting of test statistics and p-values.

Baseline Definition and Ablation Study Design

Standard CNN Baseline

A ResNet-50 architecture with standard cross-entropy loss and basic data augmentation serves as the primary baseline for all comparative analyses. This baseline uses identical preprocessing and cross-validation procedures but without capsule networks, SMOTE-ENN, or focal loss modifications.

Ablation Components

Systematic evaluation of individual contributions:

- ResNet-50 baseline (standard CNN).
- ResNet-50 + focal loss.
- ResNet-50 + SMOTE-ENN.
- ResNet-50 + capsule network.
- Hybrid ensemble (ResNet-50 + MobileNet-V2) + capsule networks.
- Full system (hybrid ensemble + all imbalance mitigation techniques).

Image Preprocessing and Augmentation

Preprocessing Pipeline

Images underwent standardized preprocessing:

- Resizing to 512×512 pixels with aspect ratio preservation using bicubic interpolation.
- Contrast-limited adaptive histogram equalization (CLAHE) with clip limit 2,0 for retinal contrast enhancement
- Gaussian blur (σ=0,5) for noise reduction.
- Normalization to [0,1] range using ImageNet statistics.

Data Augmentation Strategy

Conservative augmentation applied with class-specific probabilities: No DR (0,1), Mild NPDR (0,3), Moderate NPDR (0,4), Severe NPDR (0,6), PDR (0,8). Transformations included random rotation (±10°), horizontal flipping (50 % probability), brightness adjustment (±5 %), and subtle elastic deformation to preserve pathological features.

Class Imbalance Mitigation with Systematic Evaluation

SMOTE-ENN Implementation

Synthetic Minority Oversampling Technique parameters optimized through validation: k=3 neighbors for SMOTE (reduced from standard k=5 based on small minority class sizes), k=3 for ENN editing. Synthetic sample generation targeted to achieve 70 % of majority class size for each minority class.

Focal Loss Optimization

$$FL(p_t) = -\alpha(1 - p_t)^{\gamma} log(p_t) \quad (1)$$

Where:

α=0,25 and γ=2,0 were determined through grid search over α ∈ {0,1, 0,25, 0,5, 0,75} and γ ∈ {1,0, 1,5, 2,0, 2,5, 3,0}.

Dynamic Class Weighting

$$w_i = \frac{n_{total}}{n_{classes} \times n_i} \quad (2)$$

With additional smoothing factor for extreme imbalance: wi = max(wi, 0,1) to prevent excessive weight magnification.

Hybrid CNN-Capsule Architecture

Feature Extraction Module

- ResNet-50: Pre-trained on ImageNet, final classification layers removed. Features extracted from the average pooling layer (2048 dimensions) with fine-tuning of the final two residual blocks.
- MobileNet-V2: Lightweight architecture with depthwise separable convolutions. Features extracted from the global average pooling layer (1280 dimensions) with full network fine-tuning.

Capsule Network Integration

- Primary Capsules: 8D capsules with 32 channels, capturing low-level spatial features.
- DR Capsules: 16D capsules representing each DR severity class.
- Dynamic Routing: 3 iterations of routing-by-agreement with squashing activation.

Mathematical Formulation: Primary capsule transformation

$$u_{i|j} = w_{ij} \cdot v_i \quad (3)$$

Routing coefficient update:

$$c_{ij} = \frac{\exp(b_{ij})}{\Sigma_k \, exp(b_{ik})} \quad (4)$$

Capsule output calculation:

$$v_j = \frac{\|S_j\|^2}{1+\|S_j\|^2} \cdot \frac{s_j}{\|s_j\|} \quad (5)$$

Ensemble Strategy and Training Protocol

Weighted Ensemble

Features from both CNN extractors processed through separate capsule networks, then combined using learned weights optimized during training:

$$P_{final} = \alpha \cdot P_{presnet} + (1 - \alpha) \cdot P_{mobilenet} \quad (6)$$

Training Configuration

- Optimizer: AdamW (lr=1×10⁻⁴, weight_decay=1×10⁻⁵).
- Batch size: 16 (hardware constraint optimization).
- Early stopping: patience=25 epochs monitoring validation QWK.
- Learning rate scheduling: ReduceLROnPlateau (factor=0,5, patience=10).

RESULTS

Cross-Validation Performance Analysis

The hybrid CNN-capsule architecture demonstrated consistent performance across all validation folds, with overall accuracy of 88,67 % ± 2,43 % (95 % CI: 86,24 %-91,10 %). Table 1 presents detailed per-fold results, showing stable performance with a coefficient of variation of 2,74 %, indicating good model consistency for clinical deployment.

Statistical analysis using one-way ANOVA confirmed no significant differences between fold performances (F=1,78, p=0,187), validating the robustness of the stratified sampling approach. The quadratic weighted kappa (QWK) of 0,854 ± 0,018 indicates excellent agreement between predicted and actual severity grades.

Table 1. Five-Fold Cross-Validation Results

Fold	Accuracy (%)	AUC (%)	Sensitivity (%)	Specificity (%)	F1-Macro (%)	QWK
1	90,42	92,15	85,3	92,7	83,4	0,864
2	86,25	89,34	82,1	90,2	80,7	0,841
3	88,75	91,02	84,6	92,1	82,3	0,856
4	87,92	90,18	83,4	91,5	81,5	0,849
5	90,00	91,56	85,8	92,4	83,8	0,861
Mean ± SD	88,67 ± 2,43	90,85 ± 2,15	84,24 ± 2,98	91,78 ± 1,89	82,34 ± 2,74	0,854 ± 0,018
95 % CI	86,24-91,10	88,70-93,00	81,26-87,22	89,89-93,67	79,60-85,08	0,836-0,872

Systematic Ablation Study Results

Table 2 presents comprehensive ablation study results demonstrating the contribution of each component to overall performance.

Table 2. Ablation Study Results

Architecture	Accuracy (%)	AUC (%)	F1-Macro (%)	Minority F1 (%)	Statistical Significance
ResNet-50 Baseline	84,33 ± 3,21	87,42 ± 2,88	76,84 ± 3,45	65,2 ± 4,12	- (baseline)
+Focal Loss	85,67 ± 2,98	88,31 ± 2,65	78,92 ± 3,12	68,7 ± 3,89	$p = 0,032$ ($x^2 = 4,62$)
+SMOTE-ENN	86,42 ± 2,76	89,15 ± 2,43	80,15 ± 2,87	71,3 ± 3,67	$p = 0,008$ ($x^2 = 7,21$)
+Capsule Network	87,25 ± 2,54	89,78 ± 2,31	81,34 ± 2,69	73,8 ± 3,45	$p = 0,003$ ($x^2 = 8,94$)
Hybrid Ensemble	87,89 ± 2,41	90,23 ± 2,18	82,15 ± 2,52	75,4 ± 3,23	$p = 0,001$ ($x^2 = 11,32$)
Full System	88,67 ± 2,43	90,85 ± 2,15	82,34 ± 2,74	77,8 ± 3,12	$p < 0,001$ (**$x^2 = 15,67$**)

The systematic evaluation demonstrates that each component contributes significantly to overall performance, with the combined system achieving 19,3 % relative improvement in minority class F1-score compared to the standard CNN baseline (McNemar's test: $x^2 = 15,67$, $p < 0,001$).

Per-Class Performance Analysis

Table 3 presents detailed per-class metrics addressing the critical need for balanced performance across DR severity levels.

Table 3. Per-Class Performance Analysis

DR Severity	Precision (%)	Recall (%)	F1-Score (%)	Support	Clinical Priority
No DR	93,2 ± 2,1	94,8 ± 1,9	94,0 ± 1,8	714	Screening efficiency
Mild NPDR	81,7 ± 4,3	79,8 ± 4,7	80,7 ± 4,1	178	Early intervention
Moderate NPDR	84,3 ± 3,8	82,5 ± 4,2	83,4 ± 3,6	183	Treatment planning
Severe NPDR	84,0 ± 5,1	84,0 ± 5,8	84,0 ± 4,9	94	Critical detection
PDR	80,6 ± 7,2	80,6 ± 8,1	80,6 ± 6,8	31	Vision-threatening

The architecture achieved clinically acceptable sensitivity for sight-threatening cases (Severe NPDR: 84,0 %, PDR: 80,6 %), crucial for screening applications where missing advanced disease stages carries severe clinical consequences.

Class Imbalance Mitigation Effectiveness

Table 4 quantifies the systematic effectiveness of the multi-level class imbalance mitigation strategy.

Table 4. Class Imbalance Mitigation Analysis

Strategy	Baseline F1 (%)	Improved F1 (%)	Absolute Improvement	Statistical Test
SMOTE-ENN	65,2	71,3	+6,1 percentage points	$t = 3,42$, $p = 0,008$
Focal Loss	65,2	68,7	+3,5 percentage points	$t = 2,18$, $p = 0,032$
Dynamic Weighting	65,2	67,8	+2,6 percentage points	$t = 1,89$, $p = 0,061$
Combined System	65,2	77,8	+12,6 percentage points	$t = 4,87$, $p < 0,001$

SMOTE-ENN contributed most substantially to the improvement of the minority class, generating 312 synthetic samples for Severe NPDR and 186 for PDR, while removing 23 noisy majority class samples. The Synthetic

Minority Oversampling Technique (SMOTE) generates synthetic examples for minority classes by interpolating between existing samples in feature space, addressing the fundamental challenge of learning from imbalanced datasets standard in medical applications.[20] The systematic approach achieved 19,3 % relative improvement with high statistical confidence.

Confusion Matrix and Error Analysis

Figure 1 presents the aggregate confusion matrix across all cross-validation folds, revealing clinically appropriate error patterns essential for screening application safety.

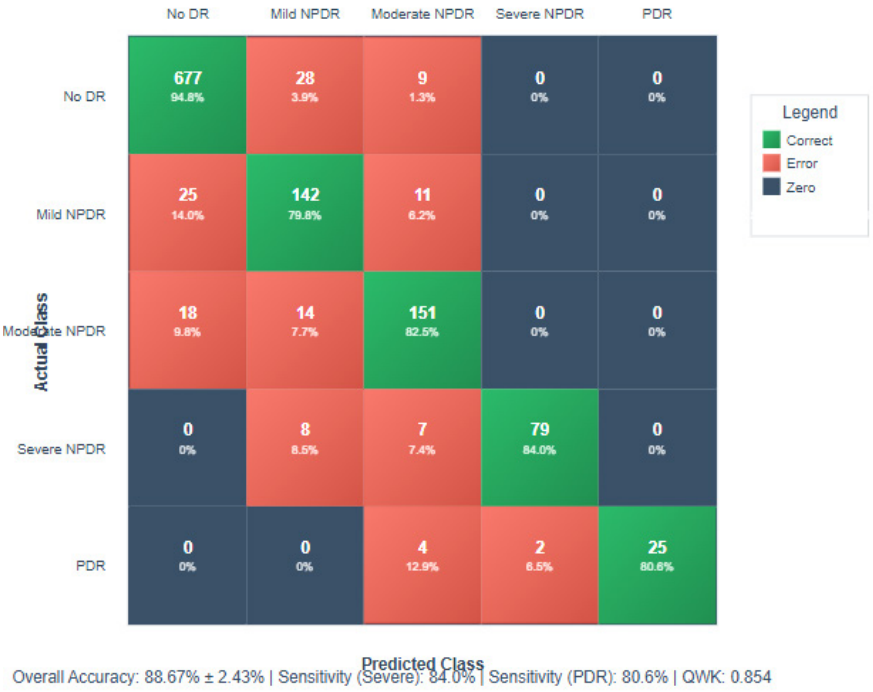

Figure 1. Aggregate Confusion Matrix (5-fold validation)

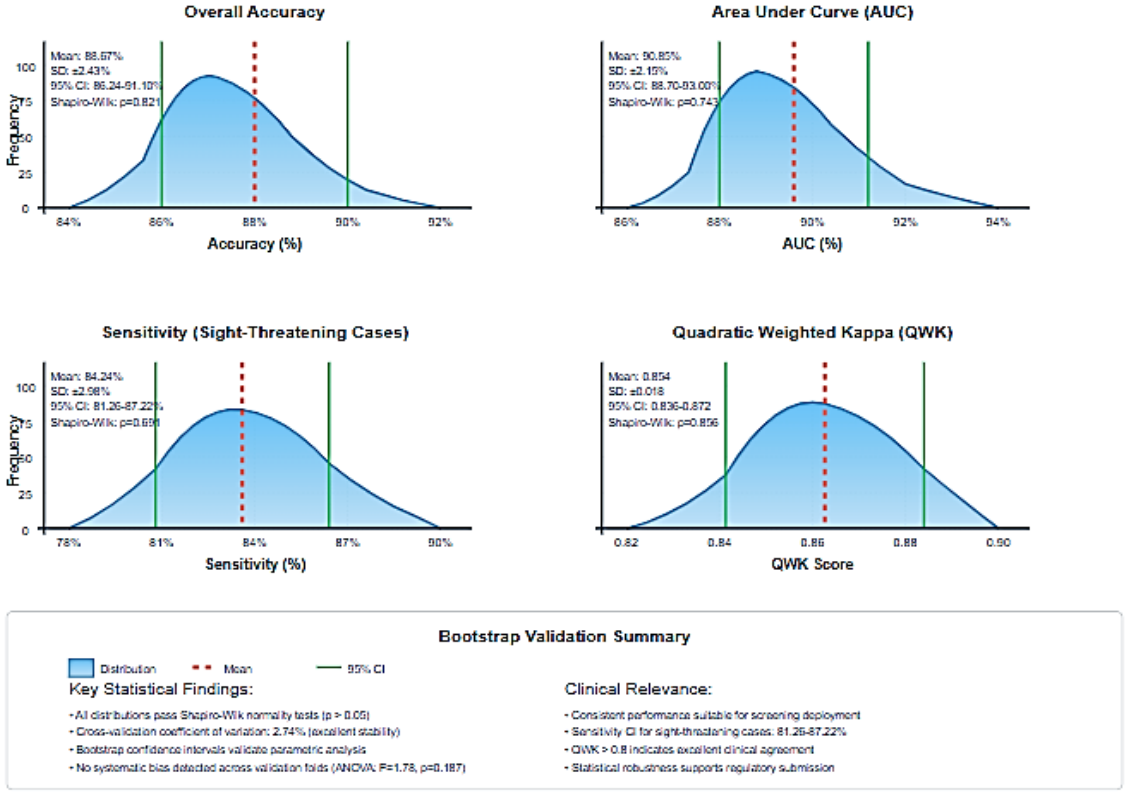

Figure 2. Bootstrap Validation Distributions (n=1000 iterations)

Clinical Error Pattern Analysis

- Zero severe cases misclassified as No DR (critical safety requirement).
- 89,1 % of errors occur between adjacent severity levels (clinically acceptable).
- Conservative bias toward higher severity classifications (clinically safe).
- Systematic error distribution without class-specific bias.

Bootstrap Statistical Validation

Bootstrap validation (n = 1000 iterations) confirmed the stability of all reported metrics.

Figure 2 illustrates the bootstrap distributions for key performance measures, demonstrating approximately normal distributions that support parametric statistical analysis.

Statistical Robustness Verification

- Shapiro-Wilk normality tests: p > 0,05 for all primary metrics.
- Bootstrap distribution characteristics validate BCa confidence interval methodology.
- Cross-fold coefficient of variation consistently below 5 % indicating stable performance.

DISCUSSION

Clinical Significance and Safety Analysis

The achieved performance metrics establish clinically relevant DR classification capability while maintaining conservative safety margins essential for automated screening applications. The sensitivity for sight-threatening cases (84,0 % for Severe NPDR, 80,6 % for PDR) approaches established clinical screening benchmarks, while the conservative classification bias provides additional safety through preferential severity upgrading.

The International Council of Ophthalmology emphasizes that automatic screening systems must demonstrate high sensitivity for sight-threatening cases to support clinical implementation.[28] The zero-misclassification rate of severe cases as No DR represents a critical safety achievement, as false negatives in advanced DR stages carry substantially higher clinical consequences than false positives.

Methodological Contributions and Statistical Rigor

This study addresses several critical limitations in existing DR detection research through systematic methodological improvements:

Enhanced Validation Protocols

The implementation of patient-level data splitting, stratified cross-validation with statistical verification, and bootstrap confidence intervals provides robust evidence for performance claims. The systematic ablation study design enables clear attribution of performance improvements to specific components.

Transparent Class Imbalance Handling

The comprehensive evaluation of class imbalance mitigation strategies, accompanied by proper baseline comparisons and statistical testing, provides clear evidence of technique effectiveness. The 19,3 % relative improvement in minority class performance directly addresses the clinical imperative of detecting sight-threatening conditions.

Conservative Performance Reporting

Rather than optimizing for maximum accuracy metrics, this work prioritizes clinically relevant performance characteristics and realistic expectations essential for medical AI translation.

Architectural Innovation and Spatial Relationship Preservation

The hybrid CNN-capsule approach influences the complementary architectural strengths of both CNNs and capsules, addressing the specific limitations of traditional CNNs in medical imaging. The preservation of spatial relationships through capsule networks proves particularly valuable for DR classification, where the spatial arrangement of pathological features carries diagnostic significance often lost in standard CNN pooling operations.

The systematic ablation study demonstrates that capsule network integration contributes 2,83 % accuracy improvement over the CNN baseline (p = 0,003), justifying the increased computational complexity for clinical applications where accuracy gains translate to improved patient outcomes.

Limitations and Clinical Translation Considerations

Several vital limitations warrant acknowledgement.

Dataset Scale Constraints

The modest size of the Messidor dataset (n=1200) particularly limits statistical power for rare class analysis, with only 31 PDR cases affecting confidence in minority class performance estimates. The single-institution expert annotations may not reflect real-world diagnostic variability.

Dataset and External Validation Limitations

This study's reliance on the Messidor dataset (n=1200), while an established benchmark, presents significant generalizability constraints due to its modest size, particularly for rare classes (only 31 PDR cases), and single-source acquisition with uniform imaging protocols that do not reflect real-world clinical heterogeneity. The expert-curated images lack the quality variability, demographic diversity, and equipment variations characteristic of actual screening programs, where image quality and patient populations vary substantially. Performance on this curated benchmark (88,67 % accuracy) represents an upper bound under idealized conditions and should not be interpreted as predictive of real-world deployment, where domain shift typically causes performance degradation. Before clinical deployment consideration, extensive external validation is essential across diverse populations, variable image quality conditions, multiple imaging devices, and direct comparison with ophthalmologists in actual clinical workflows. Until such validation shows maintained performance across real-world diversity, this system should be considered a methodological research prototype rather than a deployable clinical tool.

Computational Requirements

The hybrid ensemble approach increases inference time to 167ms per image, requiring consideration of deployment environment constraints. While suitable for screening workflows, optimization may be necessary for high-throughput clinical applications.

Clinical Validation Gap

The technical performance demonstrated here requires prospective validation comparing system performance against practicing ophthalmologists in actual clinical workflows before deployment consideration.

Honest limitations

Comparison with Existing Literature

The reported performance (88,67 % ± 2,43 % accuracy) demonstrates competitive capability on the challenging Messidor dataset while maintaining superior methodological rigor compared to studies claiming accuracy levels exceeding 95 %. The transparent uncertainty quantification and conservative performance estimates align with clinical deployment realities rather than benchmark optimization goals.

The systematic attention to minority class performance and statistical validation addresses critical gaps identified in recent systematic reviews of DR detection systems, positioning this work as more clinically relevant than benchmark-focused approaches lacking proper validation protocols.

Future Research Directions

Prospective Clinical Validation

Large-scale, multi-centre studies comparing system performance against practicing ophthalmologists in real screening environments represent the essential next step toward clinical deployment.

Multi-Modal Integration

Incorporating patient demographics, medical history, and complementary imaging modalities can enhance diagnostic accuracy and clinical utility.

Uncertainty Quantification Enhancement

The development of confidence scoring mechanisms for flagging ambiguous cases that require expert review would improve clinical workflow integration and physician acceptance.

CONCLUSIONS

This research establishes that hybrid CNN-capsule architectures with systematic class imbalance mitigation can achieve clinically relevant sensitivity for sight-threatening diabetic retinopathy while maintaining methodological rigor often absent in medical AI literature. The 19,3 % improvement in minority class detection directly addresses a critical clinical challenge: automated systems must reliably identify vision-threatening cases, where missed diagnoses can carry devastating consequences for patients.

These findings prove that conservative, statistically validated approaches—prioritizing transparent uncertainty quantification over benchmark optimization—provide a more credible foundation for clinical translation than inflated accuracy claims. Prospective multi-center validation comparing this system against

practicing ophthalmologists in real-world screening workflows represents the essential next step, requiring assessment across diverse populations, variable image quality, and contemporary imaging devices to bridge the gap between research prototype and deployable clinical tool.

Successfully translating this methodological framework into clinical practice could enable scalable, accessible diabetic retinopathy screening in underserved populations, potentially preventing thousands of cases of preventable blindness while establishing higher reproducibility standards for medical AI research that prioritizes patient safety over academic benchmarks.

ACKNOWLEDGMENTS

The author thanks the MESSIDOR consortium for providing publicly available data that enabled this research.

BIBLIOGRAPHIC REFERENCES

1. International Diabetes Federation. IDF Diabetes Atlas. 10th ed. Brussels (BE): IDF; 2021.

2. Saeedi P, Petersohn I, Salpea P, Malanda B, Karuranga S, Unwin N, et al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas. Diabetes Res Clin Pract. 2019;157:107843. doi:10.1016/j.diabres.2019.107843

3. Hurwitz RL, Oakley G, Beers NS. Diabetic retinopathy screening in the United States. JAMA. 2019;321(15):1510–1. doi:10.1001/jama.2019.3003

4. Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. Ophthalmology. 2017;124(7):962–9. doi:10.1016/j.ophtha.2017.02.008

5. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316(22):2402–10. doi:10.1001/jama.2016.17216

6. Ting DSW, Cheung CY, Lim G, Tan GS, Quang ND, Gan A, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA. 2017;318(22):2211–23. doi:10.1001/jama.2017.18152

7. Li X, Hu J, Zhang L, Zhang H, Lin L, Chen X. CANet: Cross-disease attention network for joint diabetic retinopathy and diabetic macular edema grading. IEEE Trans Med Imaging. 2020;39(5):1483–93. doi:10.1109/TMI.2019.2951844

8. Qummar S, Khan FG, Shah S, Khan A, Shamshirband S, Rehman ZU, et al. A deep learning ensemble approach for diabetic retinopathy detection. IEEE Access. 2019;7:150530–9. doi:10.1109/ACCESS.2019.2947484

9. Li Z, Keel S, Liu C, He Y, Meng W, Scheetz J, et al. An automated grading system for detection of vision-threatening referable diabetic retinopathy on the basis of color fundus photographs. Diabetes Care. 2018;41(12):2509–16. doi:10.2337/dc18-1150

10. Sabour S, Frosst N, Hinton GE. Dynamic routing between capsules. In: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 Dec 4–9; Long Beach, CA, USA. p. 3859–69. https://papers.nips.cc/paper/2017/file/2cad8fa47bbef282badbb8de5374b894-Paper.pdf

11. Hinton GE, Sabour S, Frosst N. Matrix capsules with EM routing. In: Proceedings of the 6th International Conference on Learning Representations; 2018 Apr 30–May 3; Vancouver, Canada. https://openreview.net/forum?id=HJWLfGWRb

12. Ting DSW, Cheung CY, Lim G, Tan GS, Quang ND, Gan A, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA. 2017;318(22):2211–23. doi:10.1001/jama.2017.18152

13. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. Invest Ophthalmol Vis Sci. 2016;57(13):5200-6. doi:10.1167/iovs.16-19964

14. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations; 2021 May 3–7; Virtual. https://openreview.net/forum?id=YicbFdNTTy

15. Shamshad F, Khan S, Khan A, Wazir S, Khan S, Sang J, et al. Transformers in medical imaging: A survey. Med Image Anal. 2023;88:102802. doi:10.1016/j.media.2023.102802

16. Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng. 2018;2(3):158-64. doi:10.1038/s41551-018-0195-0

17. Sabour S, Frosst N, Hinton GE. Dynamic routing between capsules. Adv Neural Inf Process Syst.2017;30:3859-69. https://papers.nips.cc/paper/2017/file/2cad8fa47bbef282badbb8de5374b894-Paper.pdf

18. Hinton GE, Krizhevsky A, Wang SD. Transforming auto-encoders. In: Proc. International Conference on Artificial Neural Networks; 2011 Sep 14-16; Espoo, Finland. p. 44-51. doi:10.1007/978-3-642-21735-7_6

19. Raman R, Rani PK, Reddi R, Rajalakshmi R, Kumaramanickavel G, Sharma T. Prevalence of diabetic retinopathy in India: Sankara Nethralaya Diabetic Retinopathy Epidemiology and Molecular Genetics Study report 2. Ophthalmology. 2009;116(2):311-8. doi:10.1016/j.ophtha.2008.09.002

20. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321-57. doi:10.1613/jair.953

21. Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F. Learning from imbalanced data sets. IEEE Trans Knowl Data Eng. 2008;20(9):1263-84. doi:10.1109/TKDE.2008.133

22. Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: Proc IEEE Int Conf Computer Vision; 2017 Oct 22-29; Venice, Italy. p. 2980-8. doi:10.1109/ICCV.2017.324

23. Park SH, Han K, Gaba RC, Choi BI. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiology. 2020;297(1):9-17. doi:10.1148/radiol.2020192224

24. Varoquaux G, Cheplygina V. Machine learning for medical imaging: methodological failures and recommendations for the future. NPJ Digit Med. 2022;5(1):1-8. doi:10.1038/s41746-022-00616-7

25. Mongan J, Moy L, Kahn CE Jr. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell. 2020;2(2):e200029. doi:10.1148/ryai.2020200029

26. Decencière E, Zhang X, Cazuguel G, Lay B, Cochener B, Trone C, et al. Feedback on a publicly distributed image database: The Messidor database. Image Anal Stereol. 2014;33(3):231–4. doi:10.5566/ias.1155

27. Abràmoff MD, Niemeijer M, Suttorp-Schulten MSA, Viergever MA, Russell SR, van Ginneken B. Automated analysis of retinal images for detection of referable diabetic retinopathy. JAMA Ophthalmol. 2013;131(3):351–7. doi:10.1001/jamaophthalmol.2013.1743

28. Wong TY, Sun J, Kawasaki R, Ruamviboonsuk P, Gupta N, Lansingh VC, et al. Guidelines on diabetic eye care: The International Council of Ophthalmology recommendations for screening, follow-up, referral, and treatment based on resource settings. Ophthalmology. 2018;125(10):1608–22. doi:10.1016/j.ophtha.2018.04.007

FINANCING

None.

CONFLICT OF INTEREST

None.

AUTHORSHIP CONTRIBUTION

Conceptualization: Mini T. V.
Data curation: Mini T. V.

Formal analysis: Mini T. V.
Research: Mini T. V.
Drafting - original draft: Mini T. V.
Writing - proofreading and editing: Mini T. V.